Bioinformática

Análisis de BRCA1

Primera entrega

Ramiro Olivera Fedi
rolivera@itba.edu.ar

Julián Antonielli
jantonielli@itba.edu.ar

Introducción

El cáncer de mama afecta a una de cada ocho mujeres durante su vida[1]. Es por esta razón que decidimos estudiar el gen `BRCA1`[2] (Breast Cancer 1), un gen supresor de tumores humano, que regula el ciclo celular y evita la proliferación incontrolada. En las mujeres portadoras de mutaciones en el gen `BRCA`1, el riesgo acumulado hasta los 70 años se estima entre 51 y 95% para cáncer de mama[3].

Se seleccionó el mRNA `NM_007294`[4] transcripto en isoforma 1 y se descargó en formato `GenBank` para el desarrollo y análisis del presente trabajo.

Implementación

El proyecto se desarrolló en Ruby con la gema Bioruby debido a la fluencia que tenemos en este lenguaje como desarrolladores. Se desarrollaron una serie de scripts, en particular, `Ex1.rb` y `Ex2.rb` cuya API e instrucciones de ejecución se exponen en el archivo `README.md`.

Ejercicio 1

El primer ejercicio consiste en leer una secuencia de nucleótidos del mRNA seleccionado (`NM_007294`) del gen `BRCA1` en formato `GenBank` y traducirlos a las posibles secuencias de aminoácidos (Dependiendo del *reading frame*), finalmente guardandolos en distintos archivos de formato `FASTA`.

La idea de la implementación es leer el archivo de `GenBank` y simplemente iterar sobre los posibles *reading frames* (De 1 a 6) traduciendo las secuencias a aminoácidos utilizando la función librería Bioruby.

La complejidad de nuestra solución se remonta básicamente al *one-liner*

```
record.to_biosequence.translate(frame)
```

que dada la representación del archivo `GenBank` de Bioruby, la transforma a una representación abstracta de una secuencia, y finalmente la traduce a la cadena de aminoácidos correspondientes con el frame indicado.

El script crea 6 nuevos archivos (uno por cada *reading frame*) con una posible secuencia de aminoácidos en el formato `FASTA`.

Ejercicio 2

Para el segundo ejercicio se tomaron las secuencias de aminoácidos generados en el ejercicio anterior en formato `FASTA` y se generó el script `Ex2.rb` que itera sobre los archivos dado el nombre del mRNA y hace consultas a `BLAST` remoto. Los resultados para

cada consulta se guardan en diferentes archivos con el mismo nombre que los generados en el ejercicio anterior pero con extensión `blas`.

El reporte generado por el `BLAST` incluye muchos resultados de interés. Si bien algunos son autodescriptivos, como el *Length of Overlapping region* que informa la cantidad de aminoácidos coincidentes entre la secuencia de la query y las encontradas en la base de datos, otros requieren de mayor atención como el valor estadístico *bit score*[6].

Para evaluar si un alineamiento dado constituye evidencia de homología, es necesario entender que el alineamiento puede darse por simple casualidad. Es por esto que es necesario cuantificar de alguna manera la significancia de la alineación con respecto al puro azar. Esto se logra a través de los puntajes, o *scores.*

Por suerte, la estadística para el cálculo de puntajes en alineamientos locales está bien estudiada y definida. Esto es particularmente cierto para alineamientos locales carentes de *huecos*.

Estos alineamientos locales consisten de un par de segmentos de igual longitud. Si los puntajes para estos no pueden mejorarse por extensión o recortado se los conoce como *HSP.*

Para analizar qué tan probable es de que surja un puntaje alto por pura casualidad, se utiliza un modelo aleatorio de secuencias. En el caso particular de secuencias suficientemente largas con longitudes `m` y `n`, la estadística de los puntajes de *HSP* se caracterizan por el parámetro `K` y `lamda`. En resumen, la cantidad de *HSPs* con puntaje de al menos S está dado por la siguiente fórmula conocida como *valor E para el puntaje S*:

$$E = Kmn\, e^{-\lambda S} \qquad (1)$$

Trabajando un poco más la fórmula anterior a través de la normalización, podemos llegar a la siguiente fórmula, con unidades fijas:

$$S' = \frac{\lambda S - \ln K}{\ln 2} \qquad (2)$$

Entonces el *valor E* correspondiente a un *bit score* dado es simplemente:

$$E = mn\, 2^{-S'} \qquad (3)$$

Los *bit score* subsumen la esencia estadística del sistema de puntuación empleado, de modo que para calcular la significancia uno necesita saber sólo el tamaño del espacio de búsqueda.

Para reconocer qué *reading frame* es el correcto utilizamos dos estrategias: A partir de los resultados del `BLAST` y utilizando la frecuencia de *codones-stop.*

Comenzamos por la segunda. La traducción del mRNA comienza cuando el ribosoma encuentra la metionina y a partir de ahí comienza a traducir. En las regiones donde los *codones-stop* son muy frecuentes, en general, no puede traducirse ningún polipéptido que tenga alguna funcionalidad específica dada la baja cantidad de aminoácidos por cadena proteica producida. En el *reading frame* 2, hay una región de alrededor de 1800 aminoácidos consecutivos a la cual se le podría llegar a asignar una funcionalidad. En contraste, los otros *reading frames* se encuentran interrumpidos en la totalidad de la secuencia, es decir, con alta frecuencia de *codones-stop* a lo largo de la secuencia.

Si nos basamos en los resultados del `BLAST` es fácil notar que el *reading frame* correcto es el segundo. Solo el hecho que el *bit score* del primer resultado de la query hecha con ese *reading frame* sea de más de 3500, mientras que el siguiente mejor con otro reading frame es de 114 es evidencia suficiente para probarlo.

Ejercicio 3

Para realizar el alineamiento múltiple se utilizó la herramienta `MUSCLE`[5], una herramienta de alineamiento de secuencias de proteínas y nucleótidos perteneciente al dominio público. `MUSCLE` es una de las herramientas de alineamiento múltiple con mejor performance de acuerdo a pruebas de referencias, con precisión y velocidad consistentemente mejores que `CLUSTALW`, con la capacidad de alinear cientos de secuencias en segundos.

Con los resultados obtenidos del `BLAST` de la secuencia de aminoácidos con el *reference frame* correcto, se seleccionaron 3 especies a comparar: *Pan paniscus, Pan troglodytes* y *Pongo abelii.*

Especie	mRNA
Homo Sapiens	NM_007294
Pan paniscus	NP_001288687
Pan troglodytes	XP_016785970
Pongo abelii	XP_003778880

Tabla 1 - mRNA seleccionado de cada especie

El repositorio de código contiene el archivo `FASTA.mix` con la concatenación de las 4

secuencias de aminoácidos, el output de la herramienta en el archivo `MUSCLE.out` y el resultado del alineamiento en `alignment.html`.

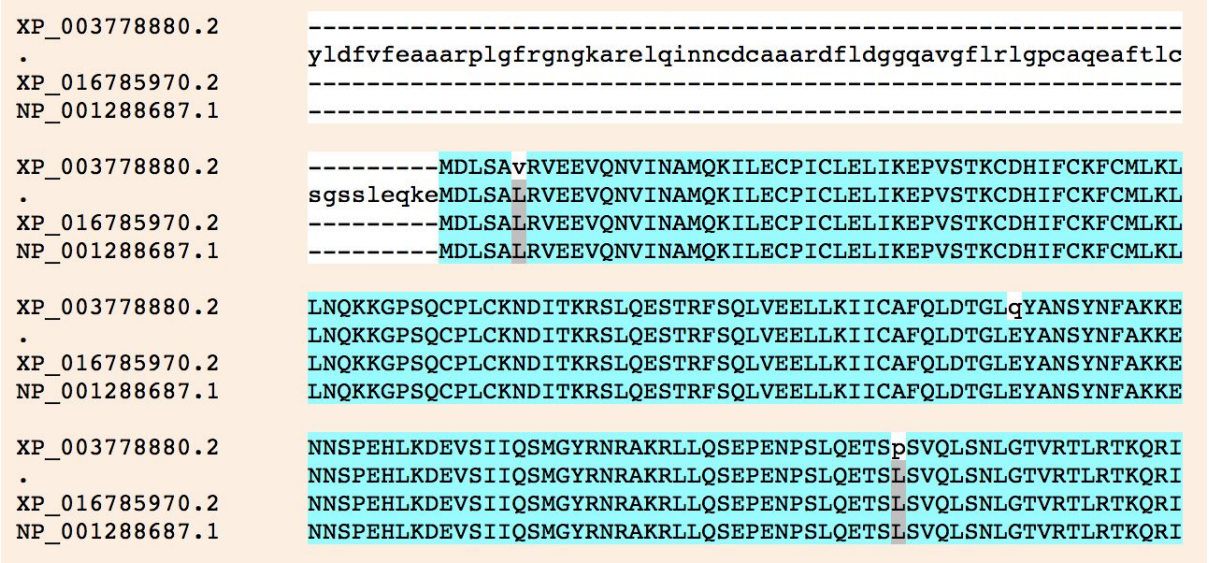

Figura 1 - Representación del alineamiento múltiple de secuencias de aminoácidos en MUSCLE

Analizando los resultados obtenidos del alineamiento podemos realizar algunas conclusiones con cierto grado de seguridad. Todas las secuencias coinciden en gran parte. De manera excepcional la cadena de aminoácidos del *Homo Sapiens* excede en el comienzo y en el final la de las especies alternativas. Podemos notar además que la secuencia conservada de la secuencia del *Homo Sapiens* (donde no hay *codones-stop*) es muy similar a la compartida entre todas las especies.

Dado que los genes en las distintas especies funcionan de igual manera, podemos concluir que la estructura de la proteína es la misma (o muy similar).

El hecho de que se trate de distintas especies podría implicar que los mismos resultados podrían obtenerse con proteínas estructuradas de distinta forma. No obstante, al tratarse de un gen regulador del ciclo celular, se trata de un mecanismo muy común conservado a lo largo de las distintas especies. Aún más en especies similares como las elegidas.

Finalmente, podemos concluir que los diferentes aminoácidos presentes en la mismas ubicaciones no alteran la estructura de la proteína, es decir, las mutaciones de estos fueron inocuas a su correcto funcionamiento.

Referencias

1. https://medlineplus.gov/spanish/breastcancer.html
2. BREAST CANCER 1 GENE; BRCA1. https://www.omim.org/entry/113705
3. Efectividad de los protocolos de prevención en mujeres portadoras de los mutaciones BRCA1/2. http://www.update-software.com/BCP/BCPGetDocument.asp?DocumentID=GCS33-7
4. Homo sapiens BRCA1, DNA repair associated (BRCA1), transcript variant 1, mRNA. https://www.ncbi.nlm.nih.gov/nuccore/NM_007294
5. Edgar RC (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput". Nucleic Acids Research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC390337
6. The Statistics of Sequence Similarity Scores. https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html